Deep Learning (CSE641/ECE555)
Quiz-2 (15 Marks) (Duration 60 min)

Name ......
Roll No. ......

Question 1-12 [1 Marks], Question 13 [3 Marks]

1. In a decoder-only transformer model that employs causal attention over a sequence of length L, what are the maximum dimensions that the attention mask matrix can have? (a) vocab_size x L (b) L $\times$ L (c) batch_size $\times$ L $\times$ L (d) L $\times$ vocab_size B. The causal attention mask matrix has dimensions L $\times$ L, where L is the context length. This creates a lower triangular matrix where each token can attend to itself and all previous tokens but not to future tokens.

2. In deep RNNs, which mathematical property of the sigmoid and tanh activation functions primarily contributes to the vanishing gradient problem?

   (a) Their output range is bounded between 0 and 1 (or -1 and 1)
   (b) Their derivatives have maximum values less than 1
   (c) Their derivatives approach zero for very large or very small inputs
   (d) All of the above

   C.

3. Which of these is not a good criterion for a good positional encoding algorithm?

   (a) It should output a common encoding for each time-step.
   (b) Distance between any two time-steps should be consistent for all sentence lengths.
   (c) It must be deterministic.
   (d) The algorithm should be able to generalize to longer sentences.

   A

4. Which of the following is the correct formula for gradient clipping?

   (a) $\hat{g} \leftarrow \hat{g}^2 + \text{threshold}$ if $\|\hat{g}\| \geq \text{threshold}$
   (b) $\hat{g} \leftarrow c$ if $\|\hat{g}\| \geq \text{threshold}$ where c is a hyper-parameter
   (c) $\hat{g} \leftarrow \frac{\text{threshold}}{\|\hat{g}\|}\hat{g}$ if $\|\hat{g}\| \geq \text{threshold}$
   (d) $\hat{g} \leftarrow \text{ReLU}(\hat{g})$ to remove negative gradients

   C

5. How many gates does a GRU (Gated Recurrent Unit) cell have?

   (a) 0; there is no gating
   (b) 1; forget gate
   (c) 2; reset and update gates
   (d) 3; reset, forget, and update gates

   C

6. Which mathematical operation is used to implement the gating mechanisms in a GRU? (a) Matrix addition (b) Matrix multiplication (Dot product) (c) Element-wise multiplication (Hadamard product) (d) Convolution (Discrete convolution) C.

7. In an LSTM cell, which computations occur inside the three gates (input, forget, and output) for the given variables: cell state c_t, hidden state h_t, and input x_t?

   (a) Forget gate: $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$, Input gate: $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$, Output gate: $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$

(b) Forget gate: $f_t = \tanh(W_f c_t + b_f)$, Input gate: $i_t = \sigma(W_i x_t + b_i)$, Output gate: $o_t = \text{ReLU}(W_o h_t + b_o)$
(c) Forget gate: $f_t = \sigma(W_f x_t + b_f)$, Input gate: $i_t = \sigma(W_i h_{t-1} + b_i)$, Output gate: $o_t = \tanh(W_o c_t + b_o)$
(d) Forget gate: $f_t = \text{softmax}(W_f[h_{t-1}, x_t] + b_f)$, Input gate: $i_t = \text{softmax}(W_i[h_{t-1}, x_t] + b_i)$, Output gate: $o_t = \text{softmax}(W_o[h_{t-1}, x_t] + b_o)$

A

8. Given the following values for an LSTM cell at time step t:

$$h_{t-1} = 0.5, \quad x_t = 0.3, \quad W_f = 1.2, \quad W_i = 0.8,$$
$$W_o = 1.5, \quad b_f = -0.1, \quad b_i = 0.2, \quad b_o = 0.05$$

Compute the output of the forget gate f_t (rounded to 2 decimal places).

(a) 0.55
(b) 0.82
(c) 0.70
(d) 0.91

C

9. What is the primary theoretical advantage of multi-head attention over single-head attention in transformer models?

(a) It reduces the computational complexity of the self-attention mechanism
(b) It allows the model to attend to information from different representation subspaces simultaneously
(c) It eliminates the need for feed-forward networks in transformer architectures
(d) It provides a more efficient alternative to recurrent neural networks

B

10. Which of the following statements about the multi-head self-attention mechanism is correct?

(a) It requires sequential processing of each head, making it much more expensive than single-head attention.
(b) It has a similar cost to single-head attention since each head operates on a lower-dimensional representation.
(c) It duplicates full-dimensional computation for each head, making it significantly more expensive.
(d) It removes the need for linear projections, reducing computational cost.

B: In multi-head self-attention, the input is projected into multiple subspaces using learned matrices, where each head operates on a lower-dimensional representation (e.g., $d_k = \frac{d_{\text{model}}}{h}$ per head). Despite having multiple heads, the overall computational cost remains similar to that of a single-head attention mechanism operating in full-dimensional space because the reduced dimensionality per head balances out the cost of having multiple heads.

11. Apart from the well-known scaled dot-product attention (SDPA) method, how else can the attention score be computed using a kernel function?

(a) $\alpha_i = K(q, k_i)$ using a similarity kernel $K(q, k)$.
(b) $\alpha_i = \frac{K(q,k_i)}{\sum_j K(q,k_j)}$ using a similarity kernel $K(q, k)$.
(c) $\alpha_i = K(q, v_i)$ instead of using keys k_i.
(d) $\alpha_i = K(q, k_i) \cdot v_i$ incorporating values directly.

B

12. In PyTorch, which function is used to reset the hidden state of an LSTM during training?

(a) lstm.reset_parameters()
(b) lstm.zero_grad()
(c) hidden_state.detach_()
(d) hidden_state = None

C

13. Consider a simple Recurrent Neural Network (RNN) for token classification with the following definitions:

- **Input at time step** t: $x_t \in \mathbb{R}^{n_x}$.
- **Hidden state**: $s_t \in \mathbb{R}^{n_h}$.
- **Output**: $y_t \in \mathbb{R}^{n_y}$.
- **Weight matrices**:
  - $W \in \mathbb{R}^{n_h \times n_x}$ (input-to-hidden weights),
  - $U \in \mathbb{R}^{n_h \times n_h}$ (hidden-to-hidden weights),
  - $V \in \mathbb{R}^{n_y \times n_h}$ (hidden-to-output weights).

The forward pass equations are:

$$s_t = tanh(Us_{t-1} + Wx_t), \tag{1}$$

$$y_t = Vs_t, \tag{2}$$

where $\phi(s)$ is an activation function, typically tanh.

The loss function is defined as:

$$L = \sum_t L_t = \sum_t \ell(y_t, \hat{y}_t). \tag{3}$$

Using Backpropagation Through Time (BPTT), derive the gradient of the loss function with respect to W.

Step 1: Compute $\frac{\partial L}{\partial s_t}$

We reuse the earlier recursive formula:

$$\frac{\partial L}{\partial s_t} = V^T \frac{\partial L_t}{\partial y_t} + U^T \frac{\partial L}{\partial s_{t+1}} \phi'(Us_t + Wx_t). \tag{4}$$

Step 2: Compute $\frac{\partial L}{\partial W}$

Since s_t depends on W as:

$$s_t = \phi(Us_{t-1} + Wx_t), \tag{5}$$

we differentiate w.r.t. W:

$$\frac{\partial s_t}{\partial W} = \phi'(Us_{t-1} + Wx_t) x_t^T. \tag{6}$$

Thus, summing over all time steps:

$$\frac{\partial L}{\partial W} = \sum_t \frac{\partial L}{\partial s_t} x_t^T. \tag{7}$$